

<http://poloclub.gatech.edu/cse6242>

CSE6242/CX4242: **Data** & **Visual** Analytics

Data Analytics Concepts

Duen Horng (Polo) Chau

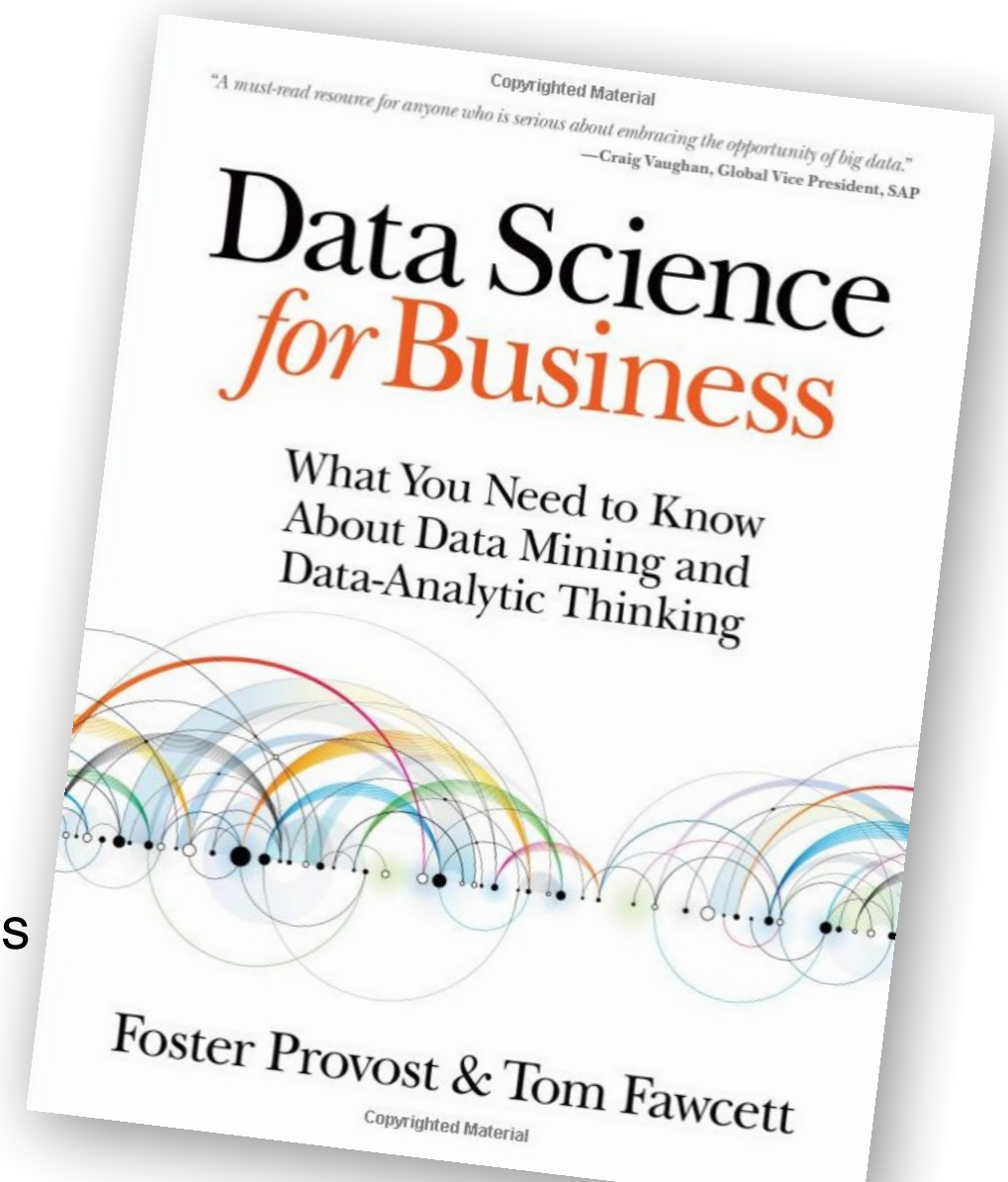
Professor, College of Computing
Associate Director, MS Analytics
Georgia Tech

Partly based on materials by Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

A critical skill in data science is the ability to decompose a data-analytics problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come in-to play.

8 concept non-mutually exclusive classes

Free for GT students



1. Classification

(or Probability Estimation)

Predict which of a (small) set of classes an entity belong to.

1. Classification

(or Probability Estimation)

Predict which of a (small) set of classes an entity belong to.

- email spam (y, n)
- sentiment analysis (+, -, neutral)
- news (politics, sports, ...)
- medical diagnosis (cancer or not)
- shirt size (s, m, l)
- cat detection
- face detection (baby, middle-aged, etc.)
- buy /not buy - commerce

2. Regression (“value estimation”)

Predict the **numerical value** of some variable for an entity.

2. Regression (“value estimation”)

Predict the **numerical value** of some variable for an entity.

- point value of wine (50-100)
- credit score
- stock prices
- relationship between price and sales
- weather
- sports and game scores

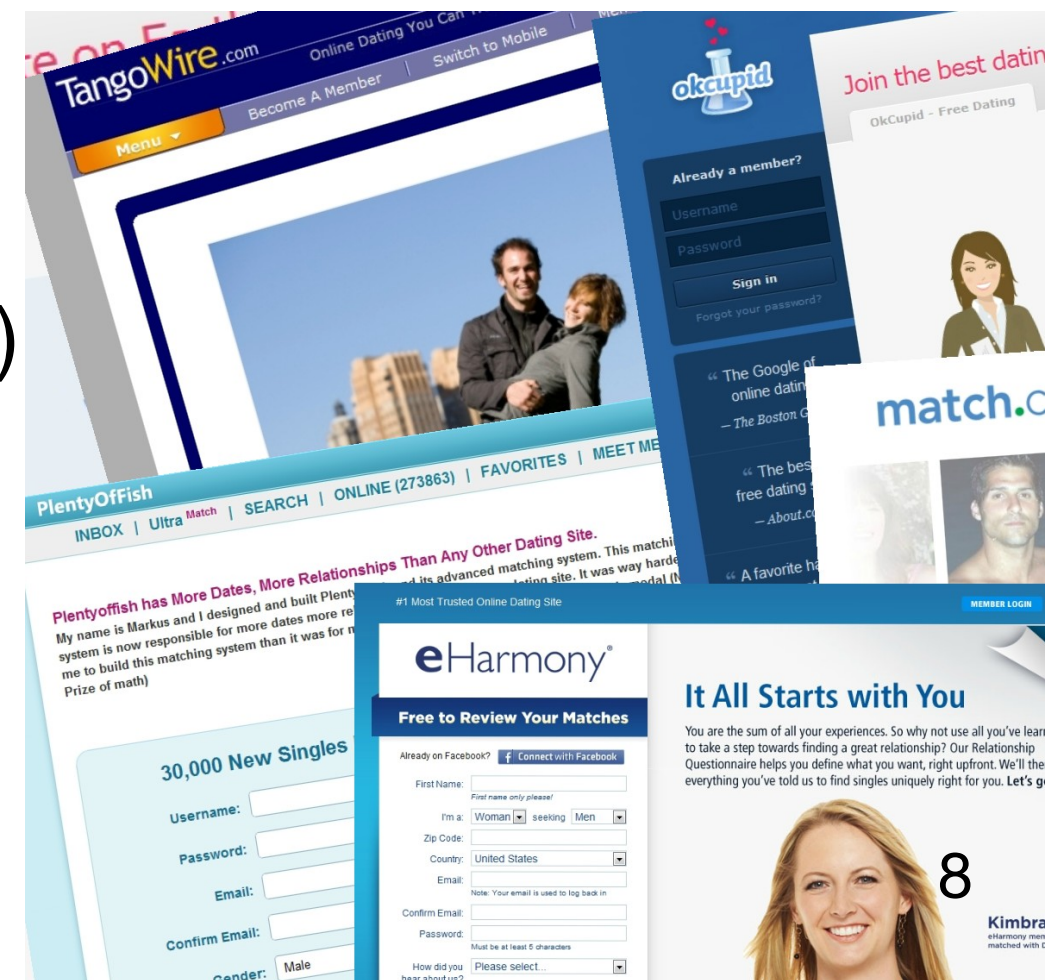
3. Similarity Matching

Find similar entities (from a large dataset) based on what we know about them.

3. Similarity Matching

Find similar entities (from a large dataset) based on what we know about them.

- find similar gene sequences (that may be repeating, or does similar things)
- online dating
- patent search
- carpool matching (find people to carpool)



4. Clustering (unsupervised learning)

Group entities together by their similarity.

(For most algorithms, user provides # of clusters)

4. Clustering (unsupervised learning)

Group entities together by their similarity.

- groupings of similar bugs in code
- topical analysis (tweets?)
- land cover: tree/road/...
- for advertising: grouping users for marketing purposes
- cluster people by accents (y'all, you all)

5. Co-occurrence grouping

(Many names: frequent itemset mining, association rule discovery, market-basket analysis)

Find associations between entities based on transactions that involve them

(e.g., bread and milk often bought together)



How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

6. Profiling / Pattern Mining / Anomaly Detection (unsupervised)

Characterize **typical** behaviors of an entity (person, computer router, etc.) so you can find **trends** and **outliers**.

- Google sign-in alert
- Computer instruction prediction
- Removing noisy data (data cleaning)
- Detect anomalies in network traffic
- **Moneyball**
- Smart security camera



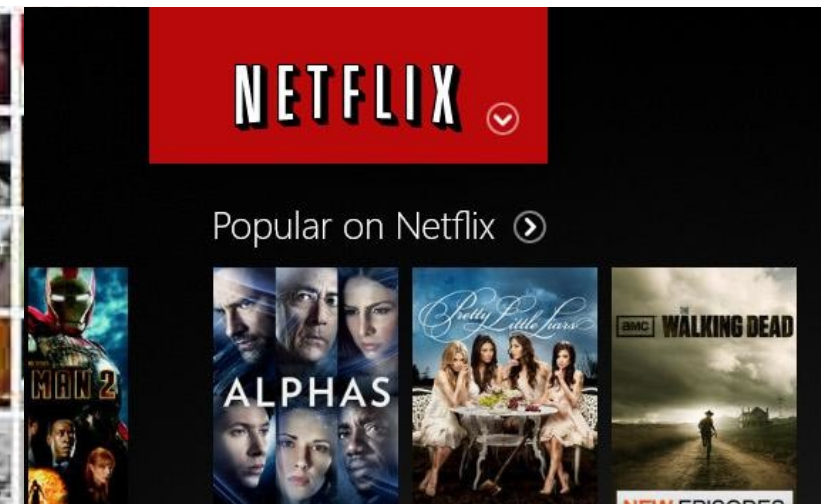
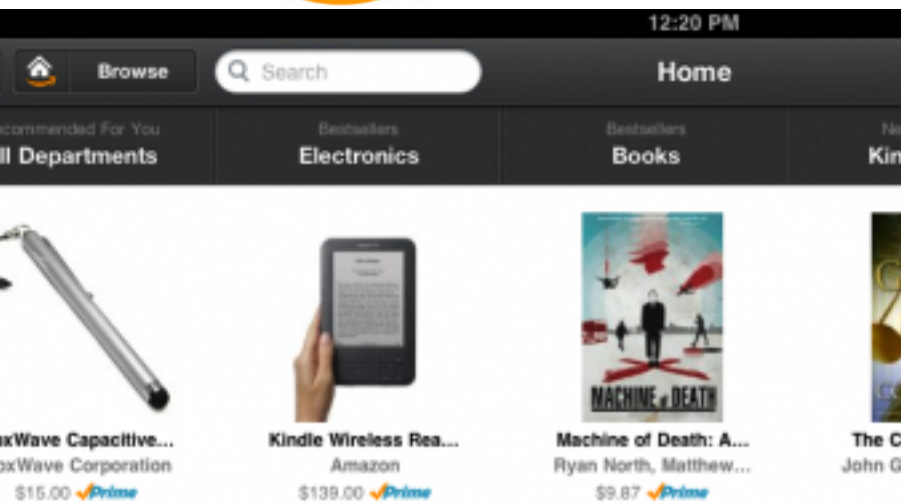
7. Link Prediction / Recommendation

Predict if two entities should be connected, and how strongly that link should be.

Linkedin/Facebook: people you may know

Amazon/Netflix.Pandora: because you like terminator...suggest other movies you may also like

amazon.com®



8. Data reduction (“dimensionality reduction”)

Shrink a large dataset into smaller one, with as little loss of information as possible

1. if you want to visualize the data (in 2D/3D)

Most popular: UMAP, T-SNE

2. faster computation/less storage

3. reduce noise

Start Thinking About Project!

- What problems do you want to solve?
- Using what **large, real** datasets?
- What techniques do you need?